Impact of Reporting Delay and Reporting Error on Cancer Incidence Rates and Trends

Limin X. Clegg, Eric J. Feuer, Douglas N. Midthune, Michael P. Fay, Benjamin F. Hankey

Background: **Cancer incidence rates and trends are a measure of the cancer burden in the general population. We studied the impact of reporting delay and reporting error on incidence rates and trends for cancers of the female breast, colorectal, lung/bronchus, prostate, and melanoma.** ***Methods:*** **Based on statistical models, we obtained reporting-adjusted (i.e., adjusted for both reporting delay and reporting error) case counts for each diagnosis year beginning in 1981 using reporting information for patients diagnosed with cancer in 1981–1998 from nine cancer registries that participate in the Surveillance, Epidemiology, and End Results (SEER) program. Joinpoint linear regression was used for trend analysis. All statistical tests are two-sided.** ***Results:*** **Initial incidence case counts (i.e., after the standard 2-year delay) accounted for only 88%–97% of the estimated final counts; it would take 4–17 years for 99% or more of the cancer cases to be reported. The percent change between reporting-adjusted and unadjusted cancer incidence rates for the 1998 diagnosis year ranged from 3% for colorectal cancers to 14% for melanoma in whites and for prostate cancer in black males. Reporting-adjusted current incidence trends for breast cancer and lung/bronchus cancer in white females showed statistically significant increases (estimated annual percent change [EAPC] = 0.6%, 95% confidence interval [CI] = 0.1% to 1.2%) and 1.2%, 95% CI = 0.7% to 1.6%, respectively), whereas trends for these cancers using unadjusted incidence rates were not statistically significantly different from zero (EAPC = 0.4%, 95% CI = –0.1% to 0.9% and 0.5%, 95% CI = –0.1% to 1.1%, respectively). Reporting-adjusted melanoma incidence rates for white males showed a statistically significant increase since 1981 (EAPC = 4.1%, 95% CI = 3.8% to 4.4%) in contrast to the unadjusted incidence rate, which was most consistent with a flat or downward trend (EAPC = –4.2%, 95% CI = –11.1% to 3.3%) after 1996.** ***Conclusions:*** **Reporting-adjusted cancer incidence rates are valuable in precisely determining current cancer incidence rates and trends and in monitoring the timeliness of data collection. Ignoring reporting delay and reporting error may produce downwardly biased cancer incidence trends, particularly in the most recent diagnosis years. [J Natl Cancer Inst 2002;94:1537–45]**

Cancer incidence rates measure the impact of cancer in the general population. The Surveillance, Epidemiology, and End Results (SEER)[1] program at the National Cancer Institute (NCI) has been actively collecting and reporting cancer incidence and survival data since 1973. The goal of a cancer registry that participates in the SEER program is to record every primary cancer in its catchment area (i.e., the geographic region for which the registry is responsible) in a timely and accurate manner, along with other information about each case, such as type of cancer, date of diagnosis, sex, race, and stage.

However, reporting delay and reporting error hamper timely and accurate case reporting. Reporting delay time refers to the time elapsed before a diagnosed cancer case is reported to the NCI. NCI's contract with registries in the SEER program specifies that the registries have up to 19 months to report cancer cases in a manner complete enough that the NCI can present the data publicly. For example, the first reports of cancer cases diagnosed in 1998 were reported to the NCI by August 2000, and data about those cases were released to the public in April 2001. Hence, the standard delay time between cancer diagnosis year and the first report of cancer incidence data to the public is about 2 years. However, as either new cancer cases are discovered or erroneous cases are detected in the existing SEER data, collected information on cases from prior diagnosis years are updated in subsequent releases of the SEER data. For example, the SEER data released in April 2001 included updated cases diagnosed from 1973 through 1997 that were reported to the NCI by August 1999 and released in April 2000.

We refer to the problem of adding new cancer cases after the standard delay time as reporting delay and deleting of erroneous cases as reporting error. Modifying collected information on reported cases may cause both reporting delay and reporting error. For example, changing a prostate cancer patient's race from white to black will cause both reporting delay and reporting error because a new cancer case for blacks is added and a previous case for whites has to be deleted. Reporting delay causes incidence rates to be underestimated and reporting error causes incidence rates to be overestimated; consequently, it is essential to adjust for both reporting delay and reporting error to obtain accurate cancer incidence rates and trends.

Reporting-adjustment of cancer incidence rates is important because of the special interest in the incidence trend (especially any change in trend) in the most recent diagnosis years. The trend in cancer incidence from the most recent diagnosis years is generally biased, however, because the most recent diagnosis year will have the largest net underreporting of cases, with smaller amounts of underreporting for less recent years. For example, Horm and Kessler *(1)* reported that lung cancer incidence rates for white males declined 4.1% from 82.7 to 79.3 cases per 100 000 person-years from 1982 to 1983 (rates age-adjusted to the 1970 U.S. standard), indicating a long-awaited downturn in lung cancer incidence consistent with an earlier decline in smoking rates. However, the current estimates for lung cancer incidence in white males are 83.8 and 82.2 in 1982

Affiliations of authors: L. X. Clegg, E. J. Feuer, M. P. Fay, B. F. Hankey (Surveillance Research Program, Division of Cancer Control and Population Sciences), D. N. Midthune (Biometry Research Group, Division of Cancer Prevention), National Cancer Institute, Bethesda, MD.

Correspondence to: Lin Clegg, Ph.D., National Cancer Institute, NIH, 6116 Executive Blvd., MSC 8316, Suite 504, Rm. 5011, Bethesda, MD 20892–8316 (e-mail: lin_clegg@nih.gov).

See "Notes" following "References."

and 1983, respectively—that is, a decline in incidence of only 1.9% *(2)*.

The idea behind modeling reporting delay and reporting error is to adjust the current case count to account for anticipated future corrections (both additions and deletions) to the data. In this study, we simultaneously model reporting delay and reporting error in SEER incidence data from diagnosis years 1981 through 1998 (using the approach of Midthune DN, Fay MP, Clegg LX, Feuer EJ: unpublished data) and examine the impact of adjustment for both reporting delay and reporting error on cancer incidence rates and trends. Hereafter, we use the term reporting-adjustment to indicate adjustment for both reporting delay and reporting error and we report age-adjusted cancer incidence rates and trends with and without reporting-adjustment. To our knowledge, this study is the first attempt to formally adjust for biases in cancer incidence rates and trends resulting from both reporting delay and reporting error. Our analysis is focused on five cancer sites: female breast, colorectal, lung/bronchus, prostate, and melanoma. We included melanoma because it has a longer reporting delay than other cancer sites, presumably because of the difficulties associated with reporting a cancer that is increasingly diagnosed in a non-hospital setting. We chose the other four cancer sites because of their high incidence rates and thus their importance in cancer control.

METHODS

Study Population and Data Source

The SEER program currently collects cancer incidence and survival information from 10 population-based cancer registries that encompass nearly 14% of the total U.S. population. This study used data from nine cancer registries that are responsible for data collection in the states of Connecticut, Hawaii, Iowa, New Mexico, and Utah and the metropolitan areas of Atlanta, Detroit, Seattle-Puget Sound, and San Francisco-Oakland. Based on 1990 census data, these nine SEER registry areas cover more than 9% of the U.S. population. These nine registries have been participating in the SEER program since 1973, except for Atlanta, which has participated in the program since 1975. This study includes all invasive primary cancers of the female breast, colorectal, lung/bronchus, prostate, and melanoma that were diagnosed in the nine SEER geographic areas between 1973 and 1998. For the reporting delay models, we used only patients diagnosed from 1981 through 1998, the years when data on reporting information were readily available. Cancer sites and morphology were coded based on the International Classification of Diseases for Oncology, second edition (ICD-O-2).

Data Structure for Delay Models

For a particular cancer, the data used for modeling were two-dimensional triangular tables of initial incidence case counts reported at the 2-year standard delay time and the additions and deletions of cases modified from the previous data submissions. Table 1 shows the summarized data for invasive melanoma. The data used in the delay models in this study correspond to a series of similar tables, one for each subgroup ($i = 1, \ldots, I$), where i indicates a subgroup with a total of I subgroups, and a subgroup is defined by every combination of the three variables that are commonly used for standard reporting (i.e., age at diagnosis, race, and sex) and the two variables for which reporting delay and error are likely to vary (i.e., registry areas and hospital versus non-hospital reporting source). A change in one of these variables for a particular case yields a reporting error in the subgroup from which the case is deleted and a reporting delay in the subgroup to which it is added.

Except at the standard delay time of 2 years, where only the initial incidence case count is displayed, there are two case counts within each cell of Table 1: the additions (i.e., the number of melanoma cases added since the previous data submission) and the deletions (i.e., the number of previously reported erroneous melanoma cases deleted). For example, 1817 melanomas diagnosed in 1981 were initially reported to the NCI in 1983. In the reporting year 1984, an additional 84 melanoma cases diagnosed in 1981 were reported, and 51 melanoma cases were deleted from the 1817 cases initially reported in 1983. In the reporting year 2000, after 19 years of delay, 10 more melanoma cases diagnosed in 1981 were added and five melanoma cases were deleted that had been previously reported erroneously as diagnosed in 1981 some time between 1983 and 1999. As of 2000, the net count for melanoma cases diagnosed in 1981 was 2048, rather than the initial 1817 (89% of 2048) reported in 1983. Over the course of 19 years, 443 new melanoma cases were added and 207 erroneous melanoma cases were deleted. It should be noted that there was a large amount of reporting error in 1993 (i.e., many cases deleted from earlier years). This irregularity in the data may reflect the possibility that some melanoma cases previously classified as "white" for those individuals of "unknown" race (because 99% of reported melanomas are from white individuals) were reclassified as "unknown" in 1993. Because we consider only melanomas for whites, we see deletions from "white" but not additions for "unknown." We postulate that the reclassification of cases based on race contributed to the large amount of reporting error in 1993, because when we combined the melanoma case counts for "white" and "unknown" so that we did not have to consider changes from "white" to "unknown" to be reporting errors, we did not see this irregular result anymore. We show the detailed data for melanoma because it has the worst case of reporting delays and reporting errors.

Modeling Delay Distributions of Cancer Case Reporting

Detailed formulation of statistical models for reporting delay and reporting error are reported elsewhere (Midthune DN, Fay MP, Clegg LX, Feuer EJ: unpublished data). Briefly, a delay distribution models the likelihood of a cancer being reported after a delay of d years ($d = 2, 3 \ldots 19$) after diagnosis. The number of cancer cases reported at each delay year (i.e., initial case count or additional cases) was assumed to follow a Poisson distribution. Cases were deleted as corrections to the data were made, and the probability of deleting cases at delay time d was modeled as a binomial distribution conditional on the net number of cases (i.e. additions minus deletions) through delay time $d - 1$.

Delay distributions were modeled as a function of covariates using a discrete-time proportional hazards model and were stratified by those variables for which the assumption of proportional hazards was not valid. For the models used in this study, diagnosis year, delay times, race (black or white), and registry were included as potential covariates. Reporting source (i.e., hospital or non-hospital) was designated as a potential stratification variable rather than a covariate because of possibly large nonproportional differences in the delay distribution. To

Table 1. Surveillance, Epidemiology, and End Results (SEER) Program incidence data for cases of invasive melanoma reported to the NCI*

Diagnosis year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998†
Reporting year																		
1983	1817																	
1984	+84 −51	1797																
1985	+31 −12	+51 −18	1762															
1986	+40 −23	+78 −21	+118 −31	1861														
1987	+25 −28	+38 −46	+36 −66	+89 −57	2117													
1988	+38 −10	+53 −12	+79 −19	+78 −10	+92 −24	2132												
1989	+40 −4	+34 −5	+55 −5	+62 −22	+92 −16	+212 −44	2334											
1990	+18 −9	+23 −13	+30 −18	+36 −13	+44 −21	+57 −12	+104 −33	2189										
1991	+43 −7	+84 −8	+102 −11	+90 −13	+89 −16	+120 −20	+168 −33	+234 −20	2530									
1992	+10 −3	+14 −2	+10 −4	+16 −7	+42 −8	+76 −14	+66 −10	+93 −19	+154 −17	2615								
1993	+12 −33	+5 −58	+15 −80	+17 −57	+27 −71	+18 −77	+20 −82	+37 −84	+37 −149	+73 −146	2721							
1994	+20 −4	+33 −3	+8 −4	+9 −3	+13 −2	+10 −3	+33 −11	+65 −5	+72 −11	+123 −14	+152 −26	2689						
1995	+34 −3	+61 −3	+89 −4	+92 −5	+76 −2	+97 −7	+100 −5	+117 −5	+165 −7	+190 −13	+251 −19	+456 −50	2960					
1996	+4 −6	+5 −1	+6 −6	+11 −7	+8 −8	+11 −6	+21 −9	+19 −9	+25 −15	+34 −14	+36 −19	+69 −24	+146 −24	3261				
1997	+5 −1	+11 −2	+5 −3	+12 −2	+17 −7	+28 −5	+21 −8	+16 −6	+32 −11	+29 −12	+31 −14	+47 −17	+86 −28	+172 −46	3514			
1998	+7 −2	+6 −4	+8 −4	+8 −5	+9 −3	+22 −2	+12 −4	+17 −3	+14 −1	+21 −10	+15 −7	+28 −10	+30 −8	+51 −15	+124 −35	3716		
1999	+22 −11	+27 −26	+31 −22	+23 −16	+40 −32	+44 −29	+43 −28	+18 −8	+31 −11	+21 −9	+24 −12	+44 −27	+58 −23	+71 −23	+106 −28	+227 −53	3860	
2000	+10 −5	+14 −6	+14 −7	+15 −10	+10 −6	+12 −8	+18 −13	+19 −16	+21 −15	+26 −12	+34 −20	+41 −31	+63 −57	+102 −85	+99 −74	+102 −64	+226 −87	3862
Initial count	1817	1797	1762	1861	2117	2132	2334	2189	2530	2615	2721	2689	2960	3261	3514	3716	3860	3862
Total +	443	537	606	558	559	707	606	635	551	517	543	685	383	396	329	329	226	—
Total −	212	228	284	227	216	227	236	175	237	230	117	159	140	169	137	117	87	—
Net count	2048	2106	2084	2192	2460	2612	2704	2649	2844	2902	3147	3215	3203	3488	3706	3928	3999	

*Number of invasive melanoma cases initially reported to the NCI for diagnosis years 1981 through 1998 and the number of cases added to (+) and deleted from (−) the previous case count in each subsequent reporting year between 1983 and 2000.

†Only the initial case count for invasive melanoma diagnosed in 1998 is presented in the reporting year of 2000 because of the standard 2-year delay time.

avoid model identifiability problems that come with modeling beyond the end of the observed data, we assumed that the number of reporting delays and reporting errors after 19 years from diagnosis was either small enough to ignore or balanced each other out.

Delay Model Selection

For each cancer site, several different models were considered. For reporting source, we either stratified by reporting source or not. The different methods of modeling the four covariates related to how these covariates were modeled. Both registry and race were either included in the model or not. Diagnosis year was 1) left out of the model, 2) modeled as a continuous covariate, or 3) modeled as a categorized covariate: 1981–1985, 1986–1990, or later than 1990. The delay times were modeled by allowing different regression parameters for the first few delay times up to some delay time k and then having one regression parameter for all delay times greater than or equal to k. The three methods of modeling the delay time have $k = 3$, $k = 5$, or $k = 10$. This approach has the advantage of smoothing the delay distribution where the data are sparse because there is less data with long delay times. The model with $k = 3$ smoothes more of the tail of the delay distribution (i.e., the portion of the distribution associated with long delay times) than the model with $k = 10$. Thus, for each cancer site except melanoma, 72 models were fit to the data as a function of reporting source (two formulations: stratified by hospital or non-hospital or not stratified), registry (two formulations: included or not), race (two formulations: included or not), diagnosis year (three formulations: included in model, modeled as a continuous covariate, or modeled as a categorized covariate), and delay time (three formulations: delay time $= 1 \ldots$, delay time $= k - 1$, and delay time $\geq k$ for $k = 3$, $k = 5$, and $k = 10$)—that is, there were 72 possible model combinations ($2 \times 2 \times 2 \times 3 \times 3$ formulations). Race was not modeled for melanoma, because we used data only from whites. Hence, only 36 models were considered for melanoma.

Each of the 72 (or 36 for melanoma) delay models were fit by the method of maximum likelihood (Midthune DN, Fay MP, Clegg LX, Feuer EJ: unpublished data). For model selection, we fit the models using incidence data from each of the annual August data submissions to the NCI between 1983 and 1998 (i.e., corresponding to diagnosis years from 1981 through 1996), and then we predicted the case counts for the 1998 diagnosis year and compared the predicted case counts with those reported in the 2000 data submission. For each cancer site, the model that minimized the sum of squared prediction errors was chosen as the default final model. However, to choose a more parsimonious model when the default model contained 24 or more parameters, we added an additional selection step in which possible competing models were selected using the following criteria: 1) the competing model had less than half of the number of parameters of the default model, and 2) the percent change between the prediction errors of the competing and the default models per extra parameter (i.e., percent change in prediction errors divided by the difference in the numbers of parameters between the two models) was less than 1%. If more than one competing model met the criteria, the model with the smallest percentage change per extra parameter was generally selected, although some ad hoc adjustments were also used. The chosen model was then refit using all data (reporting years from 1983 through 2000 for cases diagnosed from 1981 through 1998) to estimate the delay distribution. The 19-year reporting-adjusted net estimates of the cancer counts for each diagnosis year were then calculated based on the estimated delay distributions. The estimated standard errors for reporting-adjusted cancer counts accounted for variation in estimated delay probabilities. Case-reporting completeness for each diagnosis year was calculated as the ratio of the observed net count to the reporting-adjusted net counts.

Table 2. Variables included in the final delay models, by cancer site

Cancer site	Variables
Female breast	Delay = 2 years, . . . , delay = 4 years, delay ≥5 years, registry areas
Colorectal	Delay = 2 years, . . . , delay = 4 years, delay ≥5 years
Lung/bronchus	Delay = 2 years, . . . , delay = 4 years, delay ≥5 years, race
Melanoma	Delay = 2 years, delay ≥3 years, diagnosed in 1986–1990, diagnosed in 1991 or later, stratified by reporting source (hospital or non-hospital)
Prostate	Delay = 2 years, . . . , delay = 9 years, delay ≥10 years, registry areas, diagnosis year

Calculation of Incidence Rates and Trends

Age-adjusted (using the 1970 U.S. population as the standard) cancer incidence rates were calculated using cancer case counts with and without reporting-adjustment. Joinpoint linear

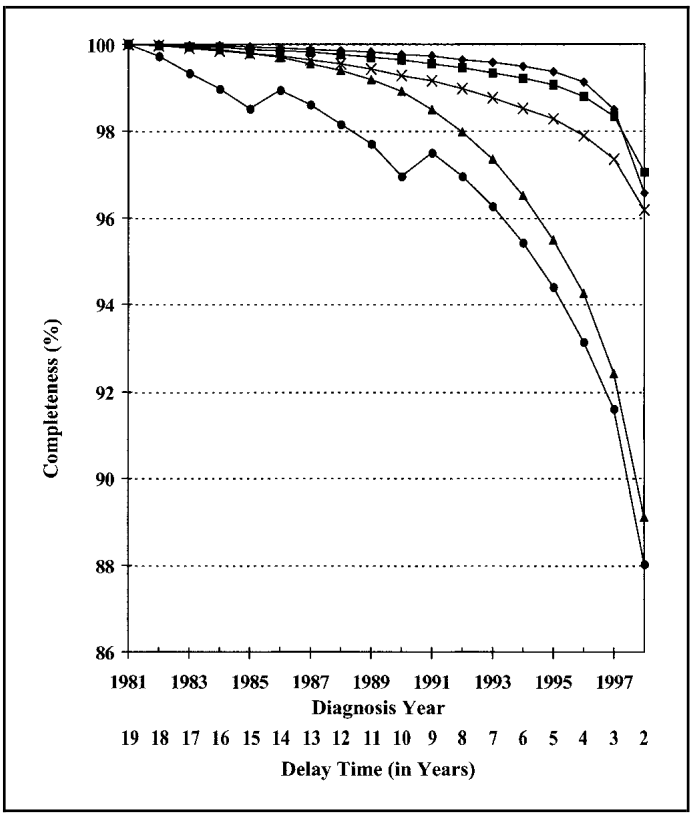

Fig. 1. Percent completeness of cancer incidence counts by diagnosis year for major cancer sites. Completeness was calculated as the ratio of case counts in each diagnosis year to the asymptotic count after 19 years, i.e., delay distribution was assumed to be complete after 19 years. The 1981 diagnosis year has 100% completeness by model assumption. Major cancer sites are as follows: female breast (×), colon/rectum (■), lung/bronchus (◆), prostate (▲), and melanoma (●). Data are from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

regression *(3,4)* was used to fit connected linear trends on a log scale—that is, where there is a constant estimated annual percentage change (EAPC) in each segment—to the 1973 through 1998 age-adjusted incidence rates both with and without reporting-adjustment. Because the delay distribution was assumed complete after 19 years, incidence rates for diagnosis years prior to 1981 were not reporting-adjusted. In joinpoint regression analyses, up to three change points (i.e., four trend line segments) were allowed, and these were modeled to fall at either whole years or midway between diagnosis years. Change points were constrained to be at least 2 years away from both the beginning and the end of the data series, and they had to be at least 2 years apart. Models were fit using weighted least squares (weighted by appropriate variances of age-adjusted incidence rates) in version 2.6 of the joinpoint regression software developed by NCI (available at http://srab.cancer.gov/joinpoint).

RESULTS

Timeliness of Cancer Case Reporting

The final delay distribution models varied by cancer site (Table 2), with only the melanoma delay model being stratified by reporting source. Based on the SEER August 2000 data submission, the 1998 incidence counts at the 2-year standard delay were about 3%–4% below the estimated numbers of cancer cases of female breast, colorectal, and lung/bronchus, according to our models (Fig. 1); more than 99% of estimated incidence counts were reported for lung/bronchus cancers diagnosed in 1996 (4-year delay), for colorectal cancers diagnosed in 1995 (5-year delay), and for female breast cancers diagnosed in 1991 (9-year delay). However, for 1998 diagnoses, only 88% of estimated melanomas and 89% of estimated prostate cancers were reported. Moreover, it would take up to 11 years for 99% of prostate cases diagnosed in 1989 to be reported and up to 17 years for 99% of melanomas diagnosed in 1983 to be reported (Fig. 1). The "jumps" in the percent completeness for melanoma in 1986 and 1991 were caused by the inclusion of diagnosis year as a discrete covariate in the delay model (diagnosis years prior to 1986, 1986–90, and after 1990), because there were incremental improvements in timeliness of case reporting in each successive time period, especially in the reporting of non-hospital cases.

Effect of Reporting-Adjustment on Cancer Incidence Rates and Trends

Figs. 2–6 depict age-adjusted cancer incidence rates and trends by site, with and without reporting-adjustment. These figures show that reporting-adjustment tended to raise cancer incidence rates in more current diagnosis years (i.e., near the end of the data series) even when it did not cause changes in the location or number of change points in the incidence trends. The percent change between reporting-adjusted and unadjusted cancer incidence rates in 1998 ranged from 3% for colorectal cancer (regardless of race or sex), to 4% for female breast cancer and lung cancer (regardless of race or sex), to 12% for prostate cancer in white males, and to 14% for melanoma in whites (regardless of sex) and prostate cancer in black males.

For female breast cancer, the reporting-adjusted incidence rates for whites (Fig. 2) in the most recent years (i.e., the last segment of the trend line between 1987 and 1998) resulted in a statistically significant increasing incidence trend in the EAPC rate of 0.6% (95% CI = 0.1% to 1.2%), whereas the trend for unadjusted incidence rates over the same time period was not statistically significantly different from zero (EAPC rate = 0.4%, 95% CI = –0.1% to 0.9%). The reporting-adjusted incidence rates also suggested an increasing trend (not statistically

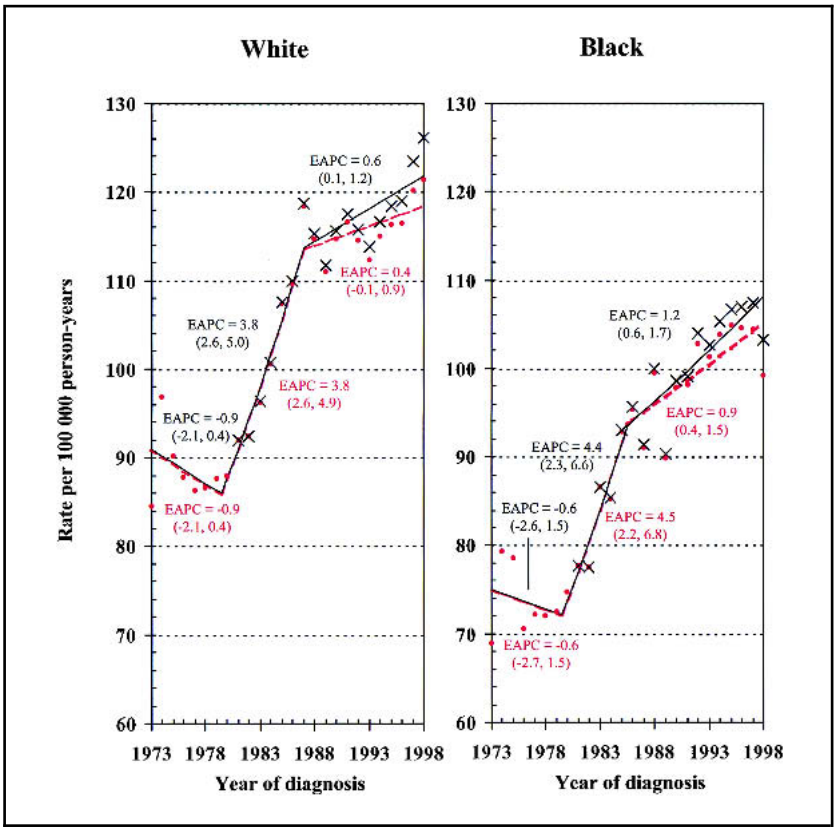

Fig. 2. Incidence and reporting-adjusted rates for female breast cancer by race. Rates are per 100 000 person-years and are age-adjusted to the 1970 U.S. standard population. (●) = incidence data and (×) = reporting-adjusted data. Regression lines are calculated using the joinpoint regression program. EAPC = estimated annual percentage change in the regression line. Numbers in parentheses are the 95% confidence intervals of the EAPC. Data are from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

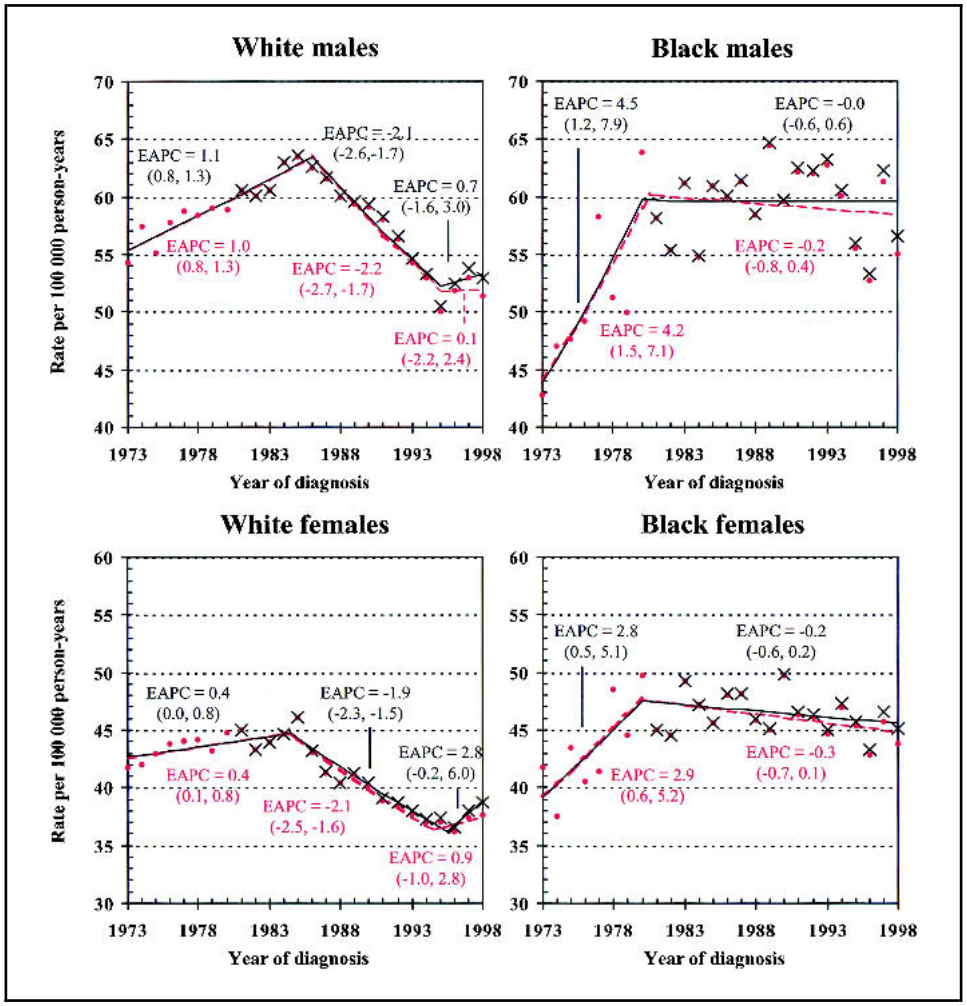

Fig. 3. Incidence and reporting-adjusted rates for colorectal cancer by race and sex. Rates are per 100 000 person-years and are age-adjusted to the 1970 U.S. standard population. (●) = incidence data and (×) = reporting-adjusted data. Regression lines are calculated using the joinpoint regression program. EAPC = estimated annual percentage change in the regression line. Numbers in parentheses are the 95% confidence intervals of the EAPC. Data are from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

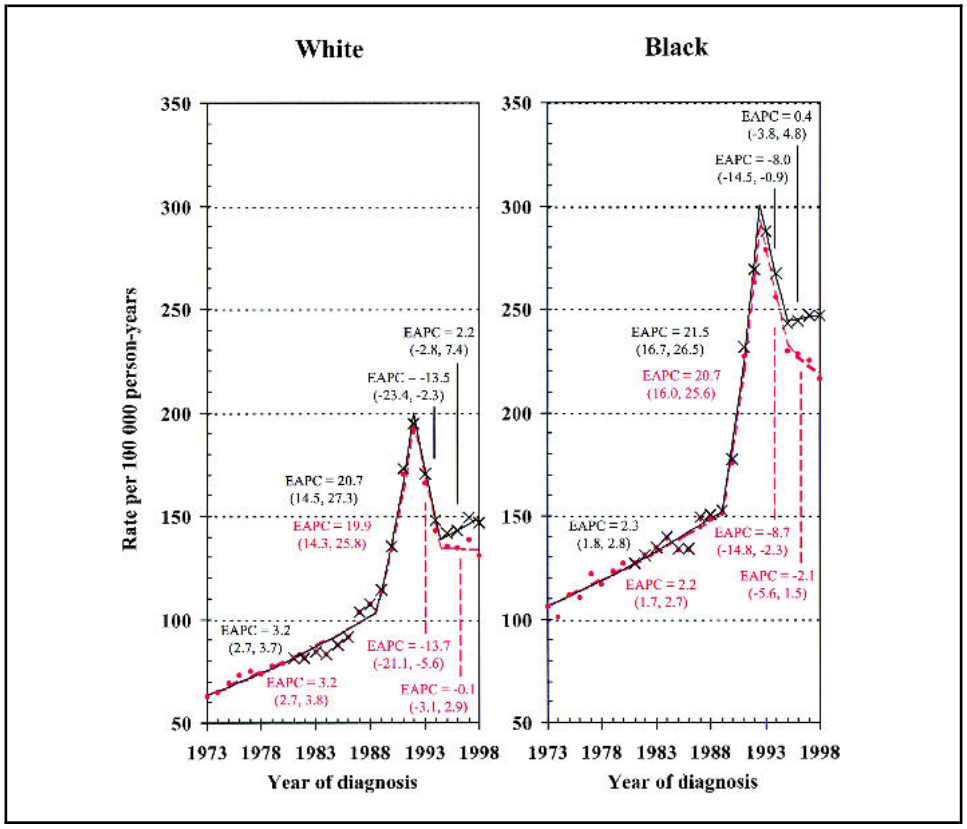

Fig. 4. Incidence and reporting-adjusted rates for prostate cancer by race. Rates are per 100 000 person-years and are age-adjusted to the 1970 U.S. standard population. (●) = incidence data and (×) = reporting-adjusted data. Regression lines are calculated using the joinpoint regression program. EAPC = estimated annual percentage change in the regression line. Numbers in parentheses are the 95% confidence intervals of the EAPC. Data are from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

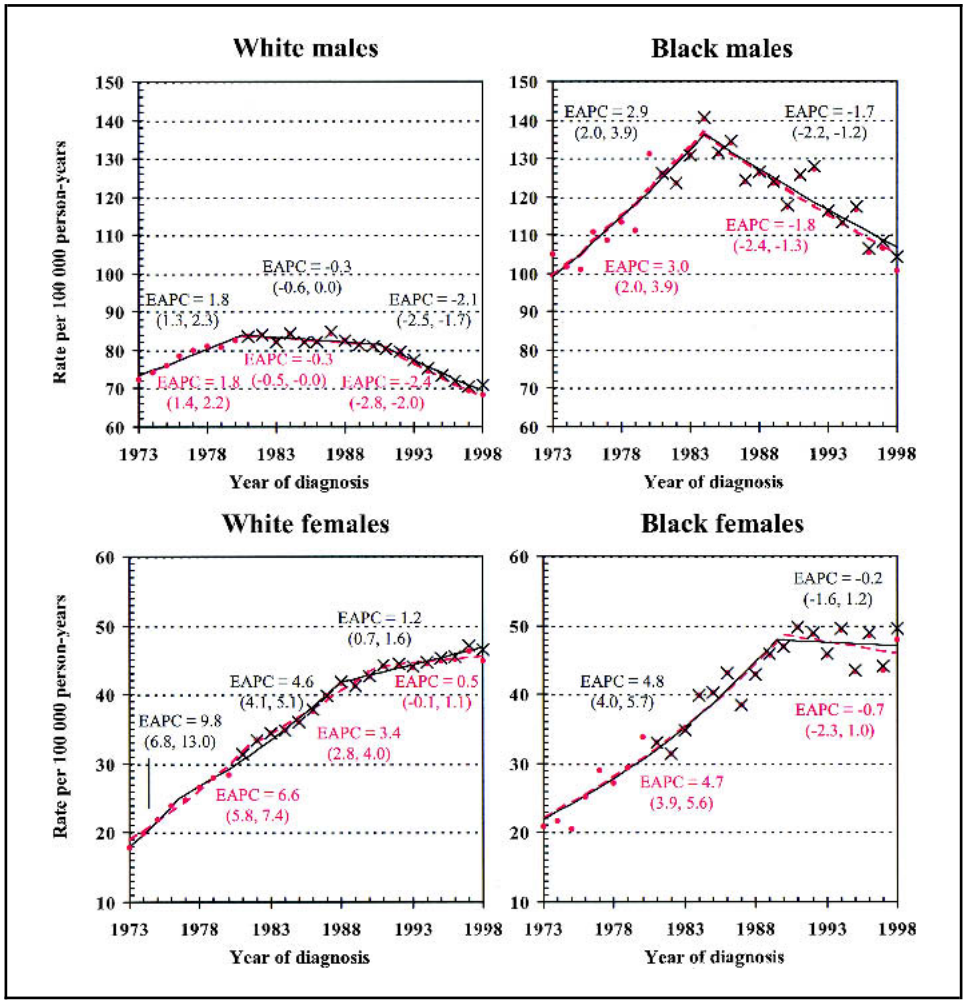

Fig. 5. Incidence and reporting-adjusted rates for lung/bronchus cancer by race and sex. Rates are per 100 000 person-years and are age-adjusted to the 1970 U.S. standard population. (●) = incidence data and (×) = reporting-adjusted data. Regression lines are calculated using the joinpoint regression program. EAPC = estimated annual percentage change in the regression line. Numbers in parentheses are the 95% confidence intervals of the EAPC. Data is from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

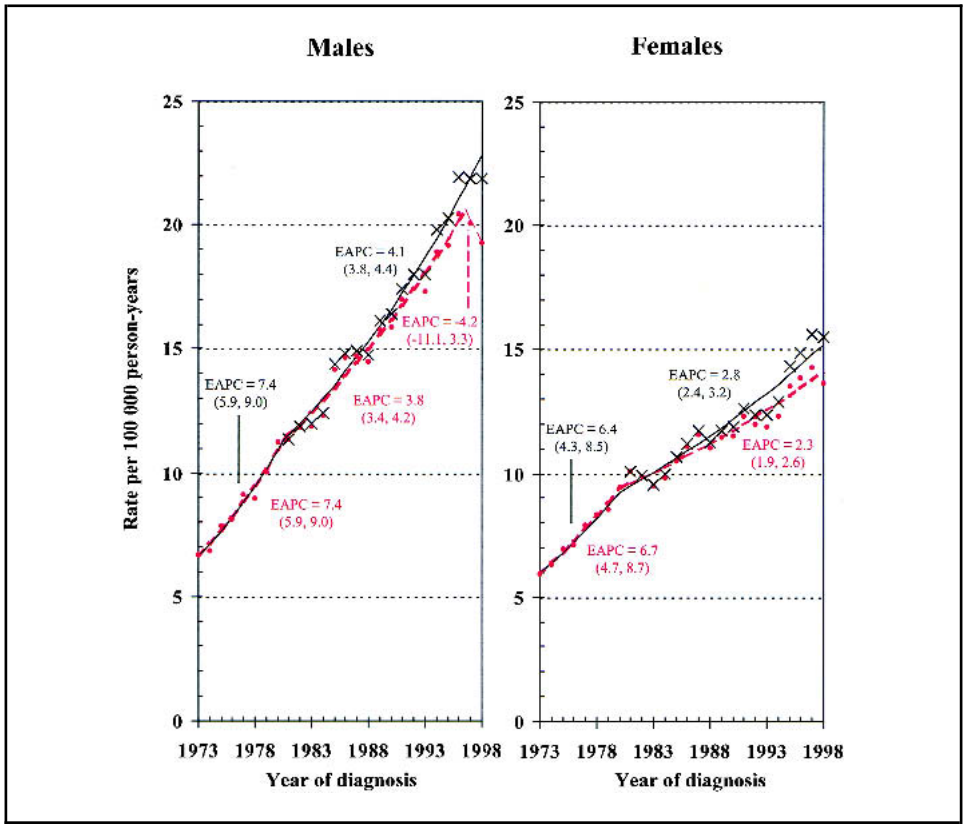

Fig. 6. Incidence and reporting-adjusted rates for melanoma in whites by sex. Rates are per 100 000 person-years and are age-adjusted to the 1970 U.S. standard population. (●) = incidence data and (×) = reporting-adjusted data. Regression lines are calculated using the joinpoint regression program. EAPC = estimated annual percentage change in the regression line. Numbers in parentheses are the 95% confidence intervals of the EAPC. Data are from the Surveillance, Epidemiology and End Results (SEER) program August 2000 submission.

significant) for colorectal cancer in white females (Fig. 3) from 1996 at an EAPC rate of 2.8% (95% CI = –0.2% to 6.0%) versus an EAPC rate of 0.9% (95% CI = –1.0% to 2.8%) for the unadjusted rates. Similarly, for prostate cancer in white males (Fig. 4) the reporting-adjusted incidence rates increased at an EAPC rate of 2.2% (95% CI = –2.8% to 7.4%), whereas the unadjusted rates showed a decrease from 1995 on with an EAPC rate of –0.1% (95% CI = –3.1% to 2.9%).

For lung/bronchus cancer in white females, reporting-adjustment caused both a shift in the change point of the trend (from 1990.5 to 1988) and an increasing EAPC rate of 1.2% (95% CI = 0.7% to 1.6%) from 1988 to 1998 rather than the flat trend (EAPC = 0.5%, 95% CI = –0.1% to 1.1%) observed for the unadjusted incidence in the most current diagnosis years (from 1991 to 1998; Fig. 5). The reporting-adjusted melanoma incidence rates for white males decreased the number of change points in the incidence trend from 3 to 2 (Fig. 6): the reporting-adjusted incidence trend continued increasing after 1980 at an EAPC rate of 4.1% (95% CI = 3.8% to 4.4%), whereas the unadjusted incidence rates indicated a flat or downward trend at an EAPC rate of –4.2% (95% CI = –11.1% to 3.3%) after the 1996 diagnosis year.

Discussion

In this article we have modeled reporting delay and reporting error to adjust the current cancer case counts for anticipated future corrections to the data. These reporting-adjusted case counts and the associated delay models are valuable in precisely determining current cancer incidence rates and trends and in monitoring the timeliness of data collection in cancer registries (Fig. 1).

Although the SEER program allows about 2 years to collect newly diagnosed cancer cases, our results show that, depending on cancer site, it would take 4–17 years for 99% or more of the cancer cases to be reported, with the incidence case counts initially reported at the 2-year delay time accounting for just 88%–97% of the estimated final incidence case counts. The low completeness percentage for prostate cancer (compared with cancers other than melanoma) might reflect an increase in diagnosis of prostate cancer in outpatient settings. Because more cases were added to the case counts after the standard 2-year delay time than were removed, the net effect of reporting-adjustment for cancer incidence case counts for the cancer sites examined was to increase cancer incidence rates in more current diagnosis years. The percent change between reporting-adjusted and unadjusted cancer incidence rates for 1998 ranged from 3% for colorectal cancer (regardless of race or sex) to 14% for melanoma (regardless of sex) and prostate cancer in black males. Thus, our results suggest that ignoring reporting delay and reporting error may result in the false impression of a recent decline in cancer incidence when the apparent decline is, in fact, caused by delayed reporting of the most recently diagnosed cases.

Reporting-adjusted incidence rates help in more accurately estimating current trends, even if the change in the rates appears to be slight. For example, there is concern over a recent rise in the incidence of breast cancer, especially for white females *(5)*. Although the reporting-adjusted incidence trends for breast cancer in white females in our study did not reveal any new joinpoints in the incidence trend, the adjusted incidence trend since 1987 was one and half times as large as the unadjusted incidence trend (an EAPC rate of 0.6% versus 0.4%). This reporting-adjusted trend is statistically significantly different from zero, whereas the observed trend (i.e., unadjusted for reporting) was not. Research efforts to explain the cause for the recent rise in breast cancer incidence rates (e.g., increases in screening mammographic examination rates, change in risk factors) are warranted.

For colorectal cancer there appear to be recent changes (albeit not statistically significant) in the incidence trends for whites occurring after 1995: the reporting-adjusted EAPC rate for the most recent incidence trend for white males is more than 10 times that of the trend for the observed data (0.7% versus 0.07%), whereas the reporting-adjusted incidence trend for females is more than three times that of the trend for the observed data (an EAPC rate of 2.8% versus 0.9%). However, one must be cautious in the interpretation of these results because the 95% CIs around these recent trends are quite large (see Fig. 3). Analyses of Medicare claims data reveal that there is a recent upswing in the rates of polypectomies (i.e., endoscopic removal of small colorectal polyps) (Brown M: personal communication), and having the best possible estimate of recent cancer incidence trends is important to help quantify the relationship between changes in medical practice and its impact on trends.

The reporting-adjusted trends for melanoma in white males revealed a continued increase in the incidence rates after 1980, whereas the trends in the observed data showed a downturn after 1997. Delayed reporting of the most recently diagnosed melanoma cases causes the false impression of a decline in the unadjusted trend.

Prostate cancer incidence trends have been under special scrutiny because the prostate-specific antigen (PSA) test-induced rise and fall in prostate cancer incidence rates from 1989 to 1995 have important implications concerning the operating characteristics of PSA testing in the community setting. If there is no overdiagnosis of prostate cancer through PSA testing, then prostate cancer incidence should return to its underlying background trend. Interestingly, our study shows that since 1995 the reporting-adjusted incidence trend for prostate cancer in white males has almost returned to its pre-PSA test-induced rise of approximately 3% per year (see Fig. 4). However, one must consider what the background incidence trend for prostate cancer would have been in the absence of the introduction of PSA screening, especially because incidental cancers detected by transurethral resection of the prostate have declined in recent years because of the introduction of drugs for the medical management of prostatic hypertrophy.

SEER cancer incidence rates are often used to validate statistical modeling on how risk factors and medical advances influence population trends. For example, the NCI is sponsoring a cooperative group of scientists known as the Cancer Intervention and Surveillance Modeling Network (CISNET). CISNET investigators are modeling the impact of cancer control interventions, such as treatment, screening, and prevention, on trends in breast, prostate, and colorectal cancer incidence and mortality and will eventually do so for lung cancer. These statistical models are to be validated using SEER cancer incidence trends. Therefore, accurate representations of current trends in cancer incidence are crucial to these modeling efforts. The NCI annual report (SEER Cancer Statistics Review; available at http://seer.cancer.gov) now includes the reporting-adjusted cancer incidence rates and

trends (in graphs similar to Figs. 2–6) for the five cancer sites we examined.

Although we modeled reporting delay and reporting error in SEER data, cancer registries (and registries for other diseases) in the United States and throughout the world could also adapt this method. A SAS/IML macro has been developed to perform these types of modeling calculations, which is available upon request from the authors. Such calculations need archived data sets from each year's data submission to make a comparison between sequential data releases to get the data in the same form as in Table 1.

Reporting delay models have previously been used in the reporting of AIDS cases *(6–9)*. However, these models have generally not explicitly modeled reporting errors and have only modeled reporting delays (Green TA: personal communication). This approach could lead to biased estimates of cancer incidence rates and trends, because reporting errors in more recent diagnosis years are less likely to have been corrected than those in earlier diagnosis years. Furthermore, not explicitly modeling reporting error would underestimate the variation of the estimates for delay distributions that results from both reporting delay and reporting error.

An important limitation of our reporting-adjustment model is that, as with all other delay models, it does not account for cases that are never reported. Underreporting could be a serious problem in monitoring cancer incidence trends if case finding at cancer registries is incomplete. Case finding is, therefore, an important aspect of data quality control for the SEER program. In addition, if reporting delays and reporting errors do not balance each other out after 19 years, the likely consequence is that unadjusted incidence rates prior to 1981 are downwardly biased as the result of both reporting delays (i.e., downward bias) and reporting errors (i.e., upward bias).

In summary, this study provides the first known research on the impact of both reporting delay and reporting error on cancer incidence rates and trends. The reporting-adjusted case counts and the associated delay models are valuable in precisely determining current cancer incidence rates and trends and in monitoring the timeliness of data collection at cancer registries.

References

(1) Horm JW, Kessler LG. Falling rates of lung cancer in men in the United States. Lancet 1986;1:425–6.

(2) Ries LA, Milton PE, Kosary CL, Hankey BF, Miller BA, Clegg LX, Edwards BK, editors. SEER cancer statistics review, 1973–1997. National Cancer Institute. NIH Publ No. 00–2789. Bethesda (MD): 2000. [Last accessed: 9/6/02.] Available at: http://seer.cancer.gov/csr/1973_1997/sections.html.

(3) Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med 2000:19:335–51.

(4) Lerman PM. Fitting segmented regression models by grid search. Appl Stat 1980:29:77–84.

(5) The NCI annual report, SEER cancer statistics review: 1973–1999. [Last accessed: 8/9/02.] Available at: http://seer.cancer.gov/csr/1973_1999/sections.html.

(6) Brookmeyer R, Damiano A. Statistical methods for short-term projections of AIDS incidence. Stat Med 1989;8:23–34.

(7) Pagano M, Tu XM, De Gruttola V, MaWhinney S. Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. Biometrics 1994;50:1203–14.

(8) Harris JE. Reporting delays and the incidence of AIDS. J Am Stat Assoc 1990;85:915–24.

(9) Green TA. Using surveillance data to monitor trends in the AIDS epidemic. Stat Med 1998;17:143–54.

Notes

[1]*Editor's Note:* SEER is a set of geographically defined, population-based, central cancer registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Registry data are submitted electronically without personal identifiers to the NCI on a biannual basis, and the NCI makes the data available to the public for scientific research.

Manuscript received February 21, 2002; revised August 8, 2002; accepted August 22, 2002.